DOCUMENT RESUME

ED 406 442 TM 026 421

AUTHOR Rennie, Kimberly M.
TITLE Understanding the Sampling Distribution: Why We Divide by n-1 To Estimate the Population Variance.
PUB DATE Jan 97
NOTE 20p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, January 23-25, 1997).
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Variance; *Estimation (Mathematics); *Sample Size; *Sampling; *Statistical Distributions; Statistical Significance
IDENTIFIERS Population; Variance (Statistical)

ABSTRACT

This paper explains the underlying assumptions of the sampling distribution and its role in significance testing. To compute statistical significance, estimates of population parameters must be obtained so that only one sampling distribution is defined. A sampling distribution is the underlying distribution of a statistic. Sampling distributions are theoretical distributions that comprise an infinite number of sample statistics from an infinite number of randomly selected samples of a specified sample size. The influence that a large sample size has on statistical significance is demonstrated through some "what if" analyses. A "what if" analysis is simply an analysis of variance summary table in which the sample size is changed to see how statistical significance is affected. A large enough sample size invariably leads to statistical significance. Researchers with large sample sizes should look for other ways to interpret their results. One such way is effect size, which is a variance accounted for statistic that can tell how much of the variability in a dependent variable can be explained by the independent variables. (Contains 2 tables, 4 figures, and 12 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 406 442

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
KIMBERLY RENNIE
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Understanding the Sampling Distribution: Why We Divide by n-1 to Estimate the Population Variance

Kimberly M. Rennie

Texas A&M University 77843-4225

Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, January, 1997.

TM026421

BEST COPY AVAILABLE

Often times, graduate students (especially those in the behavioral sciences) view statistics courses as classes in which they just have to get through. There is no desire to actually learn the material. Instead, students opt to memorize enough formulas to get a passing grade. As a result of this belief, when these graduate students take a statistics course, there is not any thinking involved. That is, they willingly accept what is being taught to them as the absolute and complete truth. Unfortunately, not all that is taught in courses or printed in books is true. Many dissertations (and research articles) contain methodological and design flaws. In fact, Thompson (1994) wrote a paper about the seven common mistakes found in dissertations. One mistake made by both graduate students and faculty alike involves the interpretation of statistical significance testing.

Significance Testing

The use of statistical significance testing in behavioral science research has been the subject of heated debate over the past two decades (Carver, 1978; Cohen, 1994; Greenwald, 1975; Thompson, 1993). More recently, the American Psychological Association (APA) has established a Task Force on Statistical Inference to consider banning the reporting of statistical significance testing in APA journals (Shea, 1996). Despite the efforts of APA and many notable researchers who argue against the improper use of statistical significance testing as the

determinant for declaring the results of a study important (Cohen, 1994; Thompson, 1989b), many researchers still rely solely on the use of statistical significance testing to claim that their findings are noteworthy (Kaminski & Good, 1996; Patel, Power, & Bhavnagri, 1996). Obviously, these researchers are not aware of the erroneous assertions that they are making. Thus, the first part of this paper will enlighten these researchers and others who are in danger of one day falling prey to the same fate by explaining, that statistical significance testing is driven in large part by sample size.

Sample size

Although, there are many reasons to argue against the use of statistical significance testing, the impact that sample size has on statistical significance testing seems to be the most salient way of demonstrating this point. "What if" analyses will be used to demonstrate how sample size directly impacts statistical significance testing (see Thompson, 1989a). A "what if" analysis is simply an ANOVA summary table in which the sample size is changed in order to see how statistical significance is affected by sample size, Tables 1 and 2 present these illustrations.

As (hopefully) all researchers know, if a sample is large enough obtaining statistically significant results is inevitable. Thompson (1996) noted that:

> statistical significance testing primarily

becomes a test of researcher endurance, because 'virtually any study can be made to show [statistically] significant results if one uses enough subjects' (Hays, 1981, p. 293) As Nunnally (1960, p. 643) noted some 35 years ago, 'If the null hypothesis is not rejected, it is usually because the N is too small. If enough data are gathered, the null hypothesis will generally be rejected.' The implication is that:

> Statistical significance testing can involve a tautological logic in which tired researchers, having collected data from hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and they know they're tired. This tautology has created considerable damage as regards the cumulation of knowledge. (Thompson, 1992, p. 436)

There is not an established method for determining the correct number of subjects that should be used in an experiment. Investigators can collect data from as few or as many subjects as they choose. Thus, conscientious researchers who collect data from a relatively large number of subjects will tend to obtain

statistically significant results regardless of the hypothesis that they are testing. This is demonstrated by the following example.

Assume that the sum-of-squares total is 100 and the data are analyzed with a one-way ANAOVA. An eccentric researcher sets out to find support for the hypothesis that people who eat apple pie have higher IQs than people who eat cherry pies. Obviously this hypothesis is pretty absurd-every self-respecting behavioral scientist knows that people who eat cherry pies are the ones with the higher IQs. However, this experimenter can obtain statistical significance for this hypothesis at a sample size of n=77 at $p<.05$, as noted in Table 2. Note that the effect size is only five percent. In this scenario, this indicates that pie preference can only account for five percent of the variance in IQ. An effect of this magnitude is not considered particularly large, according to Cohen's standards. Unfortunately, the researcher who falsely believes that statistical significance testing measures how important results are will foolishly accept and attempt to publish these findings as noteworthy.

To further see the effects of sample size on statistical significance testing, different sample sizes were entered into a "what if" equation. In all of the examples the effect size was held constant. The results clearly show that as the sample size increases, F calculated increases thereby making the probability

of statistically significant results more likely, as illustrated in Table 1.

Sampling distribution

In order to compute statistical significance, estimates of population parameters must first be obtained so that only one sampling distribution is defined (i.e., so that the sampling distribution is not statistically "indeterminate") (Thompson, 1996). Hence, the second part of this paper will explain the sampling distribution and the four properties of parameter estimates.

A sampling distribution is the underlying distribution of a statistic. Sampling distributions are theoretical distributions that are comprised of an infinite number of sample statistics taken from an infinite number of randomly selected samples of a specified sample size. For instance, if a random sample of size n=20 were taken from the population an infinite number of times, the combined means taken from all the samples would make up the sampling distribution of the mean. The ratio of the sample statistic (e.g., the mean of one sample of sample size n = 20) to the standard error of the statistic (i.e., the standard deviation of the statistic's sampling distribution) produces test statistics (e.g., *t*, *F*). These test statistics are then compared to the calculated values of the test statistics to determine if the results obtained are statistically significant.

For example, if an IQ test were given to a random sample of 100 graduate students and another to a random sample 100 of high school seniors, it is highly unlikely that the variance for the two sets of IQ scores would be the same. It is equally unlikely that either score would represent the actual population variance. Instead, these statistics would be estimates of the population variance. However, since the sample variance would tend to underestimate the actual population variance, a statistical correction (i.e., n-1) must be used in an attempt to correct for this bias. This bias and correction for this bias will be explained in more detail in the next section of the present paper.

Parameter Estimates

Parameter estimates have four properties: (a) unbiasedness, (b) consistency, (c) efficiency, and (d) sufficiency (Harnett, 1970). The properties of estimates of the population mean and estimates of the population variance will be utilized in order to explain these concepts.

Biasedness

Bias occurs when the difference between the parameter estimate and the population parameter is not equal to zero. A parameter estimate can accurately estimate, underestimate, or overestimate, the actual population parameters. In Figure 1, the parameter estimate ($\underline{X}$) perfectly estimates the actual population parameter (μ). This indicates that the parameter estimate is

equal to the actual population parameter and the estimate is unbiased (e.g., $\underline{X}=\mu$, where $\underline{X}$ is the mean estimate and μ is the population mean). Figure 2 shows an underestimate of the population parameters. In this case, the parameter estimate is less than the population parameter (e.g., $SD^2<\sigma^2$, where SD^2 is the sample variance and σ^2 is the population variance). When the parameter estimate is greater than the population parameter (e.g., $y<Y$, where y represents the parameter estimate and Y is the population parameter), this results in an overestimate of the parameter, as shown in Figure 3.

It is important to note that the mean estimate is always an unbiased estimate of the population mean and the variance estimate always underestimates the population variance. The following formula proves this fact for the mean (Harnett, 1970, p.159):

Define $\underline{X}=(1/n)(x_1+x_2+x_3+\ldots+x_n)$

$E(\underline{X})=E[(1/n)(x_1+x_2+x_3+\ldots+x_n)]$

$=1/nE[x_1+x_2+x_3+\ldots+x_n]$

$=1/n(E[x_1]+E[x_2]+E[x_3]+\ldots+x_n$

$=1/n(\mu+\mu+\mu+\ldots+\mu)$

$=1/n(n\mu)$

$\mu\underline{X}=\mu$

$\therefore E\underline{X}=\mu$

In a random sample taken from a randomly distributed population, every person in the population has an equal chance of being selected. However, every score in the population does not have an equal chance of being selected. In a randomly distributed population, extreme scores have a lower probability of being selected, as illustrated in Figure 4. In this figure, it can be seen that the extreme scores have a 1 in 16 chance of being selected versus scores at the mean which have a 1 in four chance of being selected. Thus, extreme score will tend to be underrepresented in the random sample. This results in the sample variance being lower than the variance in the population. In order to correct for this bias when calculating the variance, the SOS is divided by n-1 instead of n, which results in a larger result than when dividing b n-1.

Consistency

Consistency is the tendency of parameter estimates to become closer to the actual population parameter as the sample size increases. This occurs because it is expected that as sample size increases, the sample taken from the population becomes more representative of the population. Moreover, as sample size increases, the standard error of the statistic decreases (see Hinkle et al., 1994). Therefore, the sample statistics should become closer to the actual population values. The central limit theorem states that:

as sample size *(n)* increases, the sampling

> distribution of the mean for simple random samples of *n* cases, taken from the population with a mean of μ and a finite variance equal to σ^2, approximates a normal distribution. (Hinkle et al., 1994, p. 150)

This is also true of the variance.

Efficiency

Efficiency has to do with the credibility of parameter estimates (e.g., how reliable is the estimate?). If two estimates are unbiased, the estimate which has the smaller variance in its' sampling distribution is more efficient (see Figures 5 & 6; Mittag, 1992).

Since the mean estimate is unbiased (i.e., the mean estimate is equal to he population mean), it will also be efficient. The variance, on the other hand, is never unbiased. As a result, the variance estimate is never 100% efficient. However, as the sample size increases, the variance estimate will become more efficient.

Sufficiency

Harnett, 1970 (p. 193) defined sufficiency as an estimator that "utilizes all of the information about the population parameter that is contained in the sample data." For example, the mode, median, and range represents estimates that are not sufficient. In both the sample and the population, the mode is the most common number in the distribution, the median is the number which divides the distribution into halves having an equal

number of persons or scores in the set of ordered scores, and the range is the highest number minus the lowest number in the distribution. In all of these cases, only one or two scores are used. Meanwhile, the mean, standard deviation, and the variance are all estimates that are sufficient. The following formulas demonstrate this:

sample	population
mean = $\Sigma x/n$	$\mu = \Sigma x/N$
SD = $(x-\underline{X})/n-1$	$\sigma = (x-\mu)/N-1$
$SD^2=\sqrt{(x-\underline{X})/(n-1)}$	$\sigma^2 = \sqrt{(x-\mu)/(N-1)}$

Please note that in each of the preceding formulas every score in the distribution is utilized thereby fulfilling the requirements of being sufficient.

Conclusion

This paper explained the underlying assumptions behind the sampling distribution and its role in significance testing. Moreover, the influence that a large sample size has on statistical significance was demonstrated through "what if" analyses. A large enough sample size invariably leads to statistical significance. Researchers with large sample sizes should look for other ways to interpret their results. One such way is effect size. Effect size is a variance accounted for statistic which can tell you how much of the variability in your dependent variable can be explained by your independent variable(s).

References

Cohen, J. (1994). The earth is round ($p<.05$). American Psychologist, 45, 1304-1312.

Carver, R. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Greenwald, A. (1975). Consequences of prejudices against the null hypothesis. Psychological Bulletin, 82, 1-20.

Harnett, D. L. (1970). Introduction to statistical methods Reading, MA: Addison-Wesley Publishing Company.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1994). Applied statistics for the behavioral sciences (3rd ed.). Boston, MA: Houghton Mifflin Company.

Kaminski, R. A., & Good, R. H., III. (1996). Toward a technology for assessing basic early literacy skills. School Psychology Review, 25(2), 215-227.

Patel, N., Power, T.G., & Bhavnagri, N. P. (1996). Socialization values and practices of Indian immigrant parents: Correlates of Modernity and acculturation. Child Development, 67(2), 302-313.

Shea, C. (1996). Psychologists debate accuracy of "significance test." Chronicle of Higher Education, 42(49), A12, A16.

Thompson, B. (1989a). Asking "what if" questions about significance tests. Evaluation in Counseling and Development, 22, 66-68.

Thompson, B. (1989b). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-5.

Thompson, B. (1993). Theme issue: Statistical significance testing in contemporary practice. Journal of Experimental Education, 61(4).

Thompson, B. (1994, April). Common methodology mistakes in dissertations, revisited. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 368 771)

Table 1

Statistical Significance as Sample Size Increases

not significant-fail to reject the null

SOURCE	SOS	DF	MS	F Calc	E.S.
residual	25	1	25	2.666667	0.25
error	75	8	9.375		
TOTAL	100	9	11.11111		

not significant-fail to reject the null

SOURCE	SOS	DF	MS	F Calc	E.S.
residual	25	1	25	4.333333	0.25
error	75	13	5.769231		
TOTAL	100	14	7.142857		

significant-reject the null

SOURCE	SOS	DF	MS	F Calc	E.S.
residual	25	1	25	6	0.25
error	75	18	4.166667		
TOTAL	100	19	5.263158		

significant-reject the null

SOURCE	SOS	DF	MS	F Calc	E.S.
residual	25	1	25	7.666667	0.25
error	75	23	3.26087		
TOTAL	100	24	4.166667		

Table 2

Statistical Significance Results for Cherry Pie Example

significant-reject the null

SOURCE	SOS	DF	MS	F Calc	E.S.
residual	5	1	5	4	0.05
error	95	76	1.25		
TOTAL	100	77	1.298701		

Figure 1

Unbiased Estimate

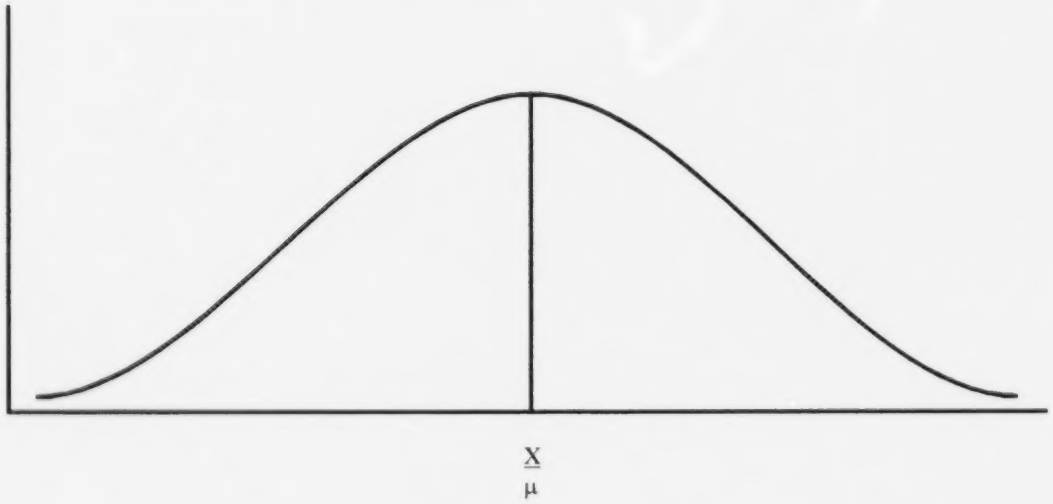


Figure 2

Underbiased Estimate

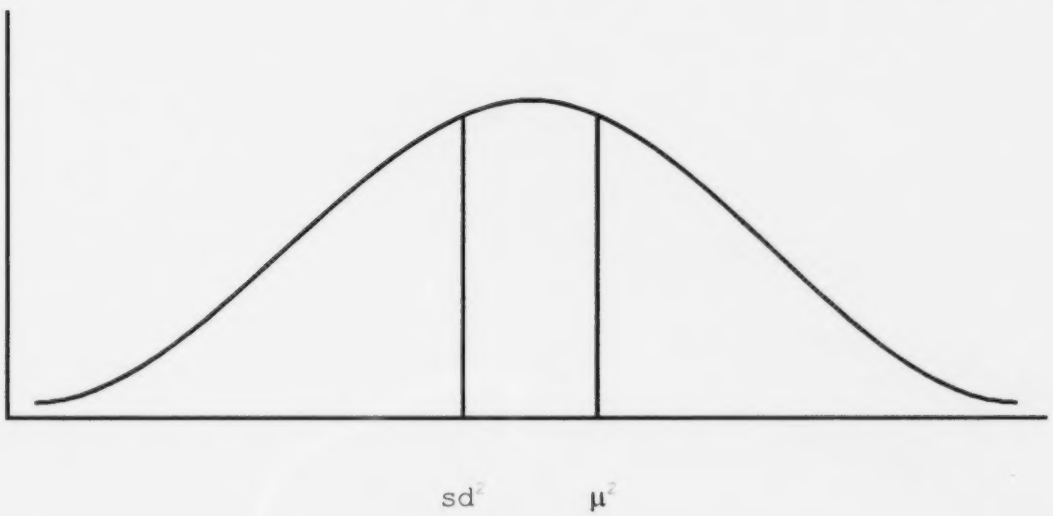
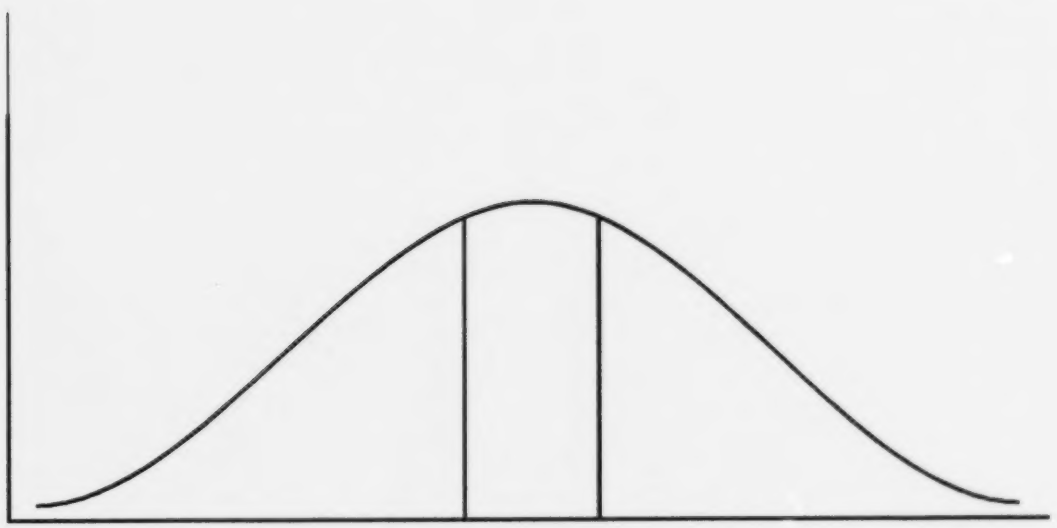


Figure 3

Overbiased Estimate



y Y

Figure 4

Probability of Selecting Extreme Scores